**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

[[1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arxiv.org)](https://arxiv.org/abs/1810.04805)

The core idea of this paper is to show to different things. First, it states the importance in using bidirectional training for language models, in comparison to both Left to right (LTR) and right to left (RTL) training. Furthermore, it also states that it is also an improvement in comparison to a model which has been trained both directions independently.

The other big part of the paper, is that it emphasizes the value of pretraining a network with the correct methods such that a user of the BERT model, only needs to finetune the model to the specified task.

Much of the article mainly consist of how the model is pretrained, and how well it performs, which is arguable pretty cool, as it performs outstandingly well, because of its bidirectional learning method.

The model is very specifically trained using sentence A as input and sentence B as another input. The purpose is to use sentence B as a tag for sentence A, such to get a form of semi unsupervised learning method. This apparently works extremely well, as Next sentence prediction (NSP) is a strong training method which has a great influence on the performance of the model.

Table

Description automatically generated

The BERT model achieves unidirectional model pre-training by using, masked language models (MLM), which a method where instead of masking all tokens on the right side of the given word in a transformer decoder architecture, the MLM chooses words by random in which should be masked. Therefore, allowing the attention of words in the future of the sentence.

Unsupervised learning is used to predict the words masked and train the network to figure out this task.

A downside for this method, is the fact that the masking is a training simulation activity, which means the model is training using a token [MASK] which will not be encountered in the finetuning training part.

To mitigate this, words that are masked, are not necessarily replaced with the [MASK] token, when a token is chosen to become masked, three things can happen to it. 80% if the time, it will be replaced by a [MASK] token. 10% of the time, it would be replaced by a random word (token, think token and word means the same thing), and the last 10% of the time it will be the unchanged token.

Not sure why this works tho?

The conclusion for this article, is the fact, that Google has successfully created a powerful bi-directional pretrained model, which only requires finetuning to perform state of the art NLP on the specific task. The emphasis is placed, on the fact, that smaller language model tasks now has the tool of a big bidirectional pretrained model, which only requires very little training time to work.